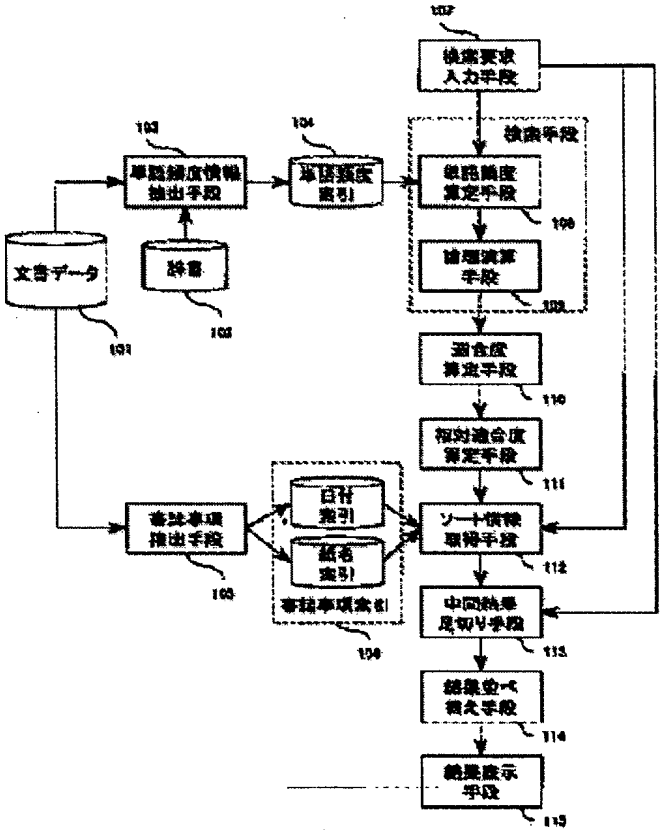
DEVICE AND METHOD FOR RETRIEVING DOCUMENT

Publication number: JP2001109766
Publication date: 2001-04-20
Inventor: INABA MITSUAKI; SUGANO YUJI
Applicant: MATSUSHITA ELECTRIC IND CO LTD
Classification:
- international: ***G06F17/30;*** **G06F17/30;** (IPC1-7): G06F17/30
- European:
Application number: JP19990288309 19991008
Priority number(s): JP19990288309 19991008

Report a data error here

Abstract of **JP2001109766**

PROBLEM TO BE SOLVED: To provide a document retrieving device capable of efficiently searching for a desired document. SOLUTION: A retrieval request character string consisting of a retrieval condition, a sort condition and the range specification of the goodness of fit to the retrieval condition is inputted from a retrieval request inputting means 107, retrieving means 108 and 109 retrieve a document meeting the retrieval condition, the goodness of fit calculating means 110 and 111 calculate the goodness of it of each document, a sort information acquiring means 112 acquires sort information, a retrieval result cutting means 113 eliminates a document that is not included in the range of the goodness of fit where the goodness of it is subjected to range specification, a retrieval result rearranging means 114 first rearranges each document according to the sort information and rearranges each document in order of the goodness of fit when the sort information is the same. It is possible to display documents while eliminating a document whose goodness of fit is not included in the range designated by a user.



(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号
特開2001－109766
（P2001－109766A）
(43)公開日　平成13年４月20日(2001.4.20)

(51)Int.Cl.[7]	識別記号	ＦＩ		テーマコード（参考）
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ　15/40	３７０Ａ	５Ｂ０７５
			３１０Ｆ	
		15/403	３７０Ｚ	
			３８０Ｅ	

審査請求　未請求　請求項の数９　ＯＬ　（全 11 頁）

(21)出願番号　特願平11－288309

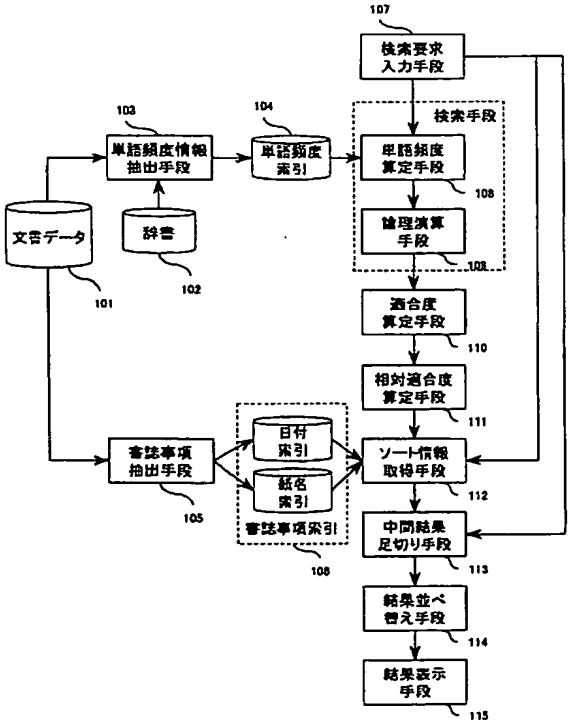
(22)出願日　平成11年10月８日(1999.10.8)

(71)出願人　000005821
松下電器産業株式会社
大阪府門真市大字門真1006番地
(72)発明者　稲葉　光昭
大阪府門真市大字門真1006番地　松下電器産業株式会社内
(72)発明者　菅野　祐司
大阪府門真市大字門真1006番地　松下電器産業株式会社内
(74)代理人　100099254
弁理士　役　昌明　（外３名）

Ｆターム(参考)　5B075 ND03 NK02 NR02 NR12 NR15
PQ29 PQ76 PQ80 PR04 PR06
QM08 UU06

(54)【発明の名称】　文書検索装置及び文書検索方法

(57)【要約】
【課題】　所望の文書を効率良く探し出すことが可能な文書検索装置を提供する。
【解決手段】　検索条件と、ソート条件と、検索条件への適合度の範囲指定とから成る検索要求文字列を検索要求入力手段107から入力し、検索手段108、109で検索条件を満たす文書を検索し、適合度算出手段110、111で各文書の適合度を算出し、ソート情報取得手段112でソート情報を取得し、検索結果足切り手段113で適合度が範囲指定された適合度範囲に入らない文書を除き、検索結果並べ替え手段114で、各文書を、まず、ソート情報で並べ替え、ソート情報が同一だった場合に適合度の順に並べ替える。適合度がユーザの指定した範囲から外れる文書を除いて表示することができる。



1

【特許請求の範囲】

【請求項１】　蓄積された文書データを検索条件にしたがって検索し、検索結果をソート条件にしたがって並べ替えて表示する文書検索装置において、

検索条件と、ソート条件と、前記検索条件に合致する度合を示す適合度の範囲指定とから成る検索要求文字列を入力する検索要求入力手段と、

前記検索条件を満たす文書を検索する検索手段と、

前記検索手段によって検索された各文書の前記適合度を算出する適合度算出手段と、

検索された前記各文書について、前記ソート条件にしたがって並べ替えを行うためのソート情報を取得するソート情報取得手段と、

検索された前記各文書から、適合度が前記範囲指定された適合度範囲に入らない文書を除く検索結果足切り手段と、

前記検索結果足切り手段から出力された前記適合度範囲に入る各文書を、まず、前記ソート情報で並べ替え、前記ソート情報が同一だった場合に前記適合度の順に並べ替える検索結果並べ替え手段と、

前記検索結果並べ替え手段によって並べ替えられた検索結果を表示する検索結果表示手段とを備えることを特徴とする文書検索装置。

【請求項２】　蓄積された文書データを検索条件にしたがって検索し、検索結果をソート条件にしたがって並べ替えて表示する文書検索装置において、

検索条件と、ソート条件と、前記検索条件に合致する度合を示す適合度の範囲指定とから成る検索要求文字列を入力する検索要求入力手段と、

前記検索条件を満たす文書を検索する検索手段と、

前記検索手段によって検索された各文書の前記適合度を算出する適合度算出手段と、

検索された前記各文書について、前記ソート条件にしたがって並べ替えを行うためのソート情報を取得するソート情報取得手段と、

検索された前記各文書の適合度を前記範囲指定された適合度範囲と比較し、前記適合度範囲に入るかどうかを示す区分けフラグを前記各文書に付与する検索結果区分け手段と、

前記検索結果区分け手段から出力された前記区分けフラグが付与された各文書を、まず、前記区分けフラグで並べ替え、前記区分けフラグの値が同一だった場合には、前記ソート情報で並べ替え、前記ソート情報が同一だった場合に前記適合度の順に並べ替える検索結果並べ替え手段と、

前記検索結果並べ替え手段によって並べ替えられた検索結果を表示する検索結果表示手段とを備えることを特徴とする文書検索装置。

【請求項３】　前記検索手段は、前記検索条件に合致する文書を検索するとともに、各文書における検索語の出現頻度を算出し、前記適合度算出手段は、前記検索手段で算出された検索語の出現頻度に基づいて各文書の前記適合度を算出することを特徴とする請求項１または請求項２に記載の文書検索装置。

2

【請求項４】　前記検索手段は、前記検索条件に合致する文書を検索するとともに、検索語の出現する文書数、及び各文書における検索語の出現頻度を算出し、前記適合度算出手段は、各文書における検索語の出現頻度と、検索語の出現文書数とに基づいて各文書の前記適合度を算出することを特徴とする請求項１または請求項２に記載の文書検索装置。

【請求項５】　前記適合度算出手段は、各文書の適合度を算定する絶対適合度算定手段と、前記絶対適合度算定手段によって算定された各文書の適合度を、それらの内の最も高い適合度に対する相対値に変換する相対適合度算定手段とを具備し、前記適合度算出手段は、各文書の適合度として前記相対値で表された相対適合度を出力し、前記検索要求入力手段は、適合度の前記範囲指定を前記相対適合度で行うことを特徴とする請求項１または２に記載の文書検索装置。

【請求項６】　蓄積された文書データを検索条件にしたがって検索し、検索結果をソート条件にしたがって並べ替えて表示する文書検索方法において、

検索条件と、ソート条件と、前記検索条件に合致する度合を示す適合度の範囲とを指定する検索要求に対して、蓄積された文書データから前記検索条件を満たす文書を検索し、検出した各文書の前記適合度を算出し、前記各文書を前記ソート条件にしたがって並べ替えるための前記各文書のソート情報を取得し、前記検索要求で指定された適合度の範囲に入らない文書を検索結果から除き、検索結果に残った各文書を、まず、前記ソート情報で並べ替え、前記ソート情報が同一だった場合に前記適合度の順に並べ替えて表示することを特徴とする文書検索方法。

【請求項７】　蓄積された文書データを検索条件にしたがって検索し、検索結果をソート条件にしたがって並べ替えて表示する文書検索方法において、

検索条件と、ソート条件と、前記検索条件に合致する度合を示す適合度の範囲とを指定する検索要求に対して、蓄積された文書データから前記検索条件を満たす文書を検索し、検出した各文書の前記適合度を算出し、前記検索要求で指定された適合度の範囲に入らない文書を検索結果から除き、検索結果に残った各文書を前記ソート条件にしたがって並べ替えるための前記各文書のソート情報を取得し、前記各文書を、まず、前記ソート情報で並べ替え、前記ソート情報が同一だった場合に前記適合度の順に並べ替えて表示することを特徴とする文書検索方法。

【請求項８】　蓄積された文書データを検索条件にしたがって検索し、検索結果をソート条件にしたがって並べ

替えて表示する文書検索方法において、
検索条件と、ソート条件と、前記検索条件に合致する度合を示す適合度の範囲とを指定する検索要求に対して、蓄積された文書データから前記検索条件を満たす文書を検索し、検出した各文書の前記適合度を算出し、前記各文書を前記ソート条件にしたがって並べ替えるための前記各文書のソート情報を取得し、前記各文書の適合度を前記検索要求で指定された適合度の範囲と比較して前記範囲に入るかどうかを示す区分けフラグを前記各文書に付与し、前記各文書を、まず、前記区分けフラグで並べ替え、前記区分けフラグの値が同一だった場合には、前記ソート情報で並べ替え、前記ソート情報が同一だった場合に前記適合度の順に並べ替えて表示することを特徴とする文書検索方法。

【請求項９】　検出した各文書の前記適合度として、前記各文書の適合度の内の最も高い適合度に対する相対適合度を算出し、前記検索要求において、適合度の範囲を前記相対適合度で指定できるようにしたことを特徴とする請求項６、請求項７または請求項８に記載の文書検索方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、検索条件にしたがって所望の文書を検索する文書検索装置と文書検索方法に関し、特に、各文書が検索条件に合致する度合と、各文書に付随する書誌事項、例えば新聞記事ならば日付の新しい順などの組み合わせによって検索結果を並べ替えて表示できるようにしたものである。

【０００２】

【従来の技術】近年、文書中における検索語の出現頻度等に基づいて、文書と検索条件との適合度を求め、その高い順に結果を並び替えて表示する、文書ランキングの手法が注目されてきている。さらに、文書に付随する書誌事項、例えば新聞記事であれば日付をソート条件として指定し、日付の新しい記事から優先して表示するが、同一日付の記事については検索条件との適合度の高い順に表示するといった、柔軟な検索が実現されてきている。

【０００３】従来の文書検索装置は、図１３に示すように、検索対象となる新聞記事の文書データ1301から辞書1302に載る単語の単語頻度情報を抽出し、単語頻度索引1304に格納する単語頻度情報抽出手段1303と、文書データ1301から日付・紙名コードといった書誌事項の情報を取り出し、書誌事項索引1306に格納する書誌事項抽出手段1305と、ユーザが検索条件及びソート条件からなる検索要求文字列を入力するための検索要求入力手段1307と、単語頻度索引1304を調べて検索条件に含まれる検索語の文書中での出現頻度を求める単語頻度算定手段1308と、レコード集合間の論理演算を行う論理演算手段1309と、検索条件と各レコードとの適合度を算出する適合度

算定手段1310と、ソート条件に指定された並べ替えのための書誌情報を取得するソート情報取得手段1311と、書誌情報と適合度とによって検索結果のレコードリストを並べ替える結果並べ替え手段1312と、検索結果を表示する結果表示手段1313とを備えている。

【０００４】なお、単語頻度索引1304には、単語頻度情報抽出手段1303の抽出動作により、検索対象文書中の辞書単語の出現頻度が格納される。

【０００５】図１４は、従来の文書検索装置における検索の処理手順を示すフローチャートである。文書データ1301は、レコード区切り文字で区切られた複数のレコード（文書）から成り、各レコードは、フィールド区切り文字で区切られた複数のフィールドから成っている。図３は文書データの具体例を示しており、フィールド区切り文字が「＾Ｆ」、レコード区切り文字が「＾Ｒ」で、紙名コード、日付、記事本文という３つのフィールドから成る新聞記事データである。

【０００６】単語頻度情報抽出手段1303は、予め文書データ1301を走査し、辞書1302に登録されている単語が各レコードの記事本文フィールドに何回出現しているかをカウントし、当該単語が出現しているレコード数及び総レコード数とともに、単語頻度索引1304に格納する。

【０００７】また、書誌事項抽出手段1305は、予め文書データ1301を走査し、各レコードの書誌事項フィールドの内容を書誌事項索引1306に格納する。

【０００８】まず、
ステップ1401：ユーザは検索要求入力手段1307により、検索要求文字列を入力する。検索要求文字列は検索条件、ソート条件の２つの部分からなる。図１５は検索要求文字列の具体例を示しており、「松下 ＡＮＤ 新製品」の部分は検索条件で、「松下」と「新製品」という２つの検索語をともに記事本文に含むような記事を検索することを意味し、「@ＨＩＤＵＫＥ @ＳＨＩＭＥＩ」の部分はソート条件で、検索結果を日付の新しい順で並べ、同じ日付なら紙名コードの小さい順で並べるということを意味している。日付、紙名コードがどちらも同じ場合は適合度の順に並べる。

【０００９】ステップ1402：単語頻度算定手段1308は、全ての検索語を対象として、
ステップ1403：単語頻度索引1304を参照し、検索要求入力手段1307によって入力された検索条件に含まれる検索語について、当該単語が記事本文に出現するレコード数と各レコードの内部番号、各レコードにおける当該単語の出現頻度及び総レコード数を算出する。

【００１０】ステップ1404：論理演算手段1309は、単語頻度算定手段1308の出力したレコード集合間の論理演算を行う。

【００１１】ステップ1405：適合度算定手段1310は、全ての検索結果レコードを対象として、
ステップ1406：論理演算手段1309の出力した各レコード

5

について、検索条件との適合度（Ｒｅｌ）を、たとえば

$$Rel = \Sigma (TFi \cdot IDFi)$$

$$IDFi = 1 - \log_2 (DFi/ND)$$

ただし、ＴＦｉは検索語Ｗｉのレコード内出現頻度、ＤＦｉは語Ｗｉの出現するレコード数、ＮＤは総レコード数を表す。

【００１２】なお、適合度の算出方法は（数１）に限らない。

【００１３】ステップ1407：ソート情報取得手段1311は、書誌事項索引1306を参照し、適合度算定手段1310の出力した各レコードの、検索要求入力手段1307から入力されたソート条件に対応する書誌事項の値をソート情報として取得する。

【００１４】図６はソート情報取得手段1311の出力内容例を示しており、日付と紙面コードの値をソート情報として取得している。

【００１５】ステップ1408：結果並べ替え手段1312は、ソート情報として取得した複数の書誌事項をソートキーとして、ソート情報取得手段1311の出力を並べ替えて出力する。このとき、すべての書誌事項の値が同じレコードがあった場合には適合度の大きい順に並べ替える。

【００１６】図１６は、結果並べ替え手段1312の出力内容の例である。

【００１７】ステップ1409：結果表示手段1313は、結果並べ替え手段1312の出力を整形してユーザに提示する。

【００１８】

【発明が解決しようとする課題】しかし、従来の構成では、並べ替えのキーとして、適合度よりも、ソート条件に指定した書誌事項の値などが優先されるために、適合度の低い文書が上位に、適合度の高い文書が下位にランクされてしまうことがあり、所望の文書を効率良く探し出すことができないという問題点があった。

【００１９】たとえば、図８において最下位にランクされている文書（レコード内部番号10）がこれに当たる。

【００２０】本発明は、こうした従来技術の課題を解決するものであり、ソート条件に指定された書誌事項の値を並べ替えのキーとして重要視しながらも、ユーザが適合度の範囲を限定することができ、指定した適合度範囲に入らない文書を結果から除いたり、より下位にランクすることで、所望の文書を効率良く探し出すことが可能な文書検索装置を提供し、また、その文書検索方法を提供することを目的としている。

【００２１】

【課題を解決するための手段】そこで、本発明の文書検索装置では、検索要求文字列として、検索条件、ソート条件に加え、適合度範囲指定を入力する検索要求入力手段と、適合度が指定された適合度範囲に入らない文書を検索結果から除く検索結果足切り手段とを設けている。

【００２２】また、検索要求文字列として、検索条件、ソート条件に加え、適合度範囲指定を入力する検索要求入力手段と、文書の適合度が、指定された適合度範囲に該当するかどうかにより、異なる区分けフラグを付与する中間結果区分け手段とを設けている。

6

（数１）によって算出する。

（Σはｉについて加算）　　（数１）

【００２３】また、本発明の文書検索方法では、検索条件とソート条件と適合度の範囲とを指定する検索要求に対して、蓄積された文書データから検索条件を満たす文書を検索し、検索した各文書の適合度を算出し、各文書のソート情報を取得し、検索要求で指定された適合度の範囲に入らない文書を検索結果から除き、検索結果に残った各文書を、まず、ソート情報で並べ替え、ソート情報が同一だった場合に適合度の順に並べ替えて表示するようにしている。

【００２４】また、この各文書のソート情報を取得する手順と、検索要求で指定された適合度の範囲に入らない文書を検索結果から除く手順とを入れ替えている。

【００２５】また、検索条件と、ソート条件と、検索条件に合致する度合を示す適合度の範囲とを指定する検索要求に対して、蓄積された文書データから検索条件を満たす文書を検索し、検出した各文書の適合度を算出し、各文書をソート条件にしたがって並べ替えるための各文書のソート情報を取得し、各文書の適合度を検索要求で指定された適合度の範囲と比較して、その範囲に入るかどうかを示す区分けフラグを各文書に付与し、各文書を、まず、区分けフラグで並べ替え、区分けフラグの値が同一だった場合には、ソート情報で並べ替え、ソート情報が同一だった場合に適合度の順に並べ替えて表示するようにしている。

【００２６】そのため、適合度がユーザの指定した範囲から外れる文書を、検索結果から除いたり、より下位にランク付けすることができ、ソート条件を指定した場合の、適合度の低い文書が上位に、適合度の高い文書が下位にランクされてしまうという問題を回避することができる。

【００２７】

【発明の実施の形態】以下、本発明の実施の形態について、図を参照しながら説明する。

【００２８】（第１の実施の形態）図１は本発明の第１の実施形態における文書検索装置の構成を示したブロック図である。

【００２９】この装置は、従来の装置（図１３）と同様に、検索対象となる新聞記事の文書データ101から辞書102に載る単語の単語頻度情報を抽出して単語頻度索引104に格納する単語頻度情報抽出手段103、文書データ101から日付・紙名コードといった書誌事項の情報を取り出して書誌事項索引106に格納する書誌事項抽出手段105、検索要求入力手段107、単語頻度算定手段108、論理演算手段109、適合度算定手段110、ソート情報取得手段11

2、結果並べ替え手段114、及び、結果表示手段115を備えるとともに、適合度算定手段110によって算定された各レコードの適合度を最大値に対する相対値へ変換してソート情報取得手段112に出力する相対適合度算定手段111と、ソート情報取得手段112から出力された検索結果から適合度の値が指定した適合度範囲に入らないレコードを除く中間結果足切り手段113とを備えている。

【００３０】図２のフローチャートは、第１の実施形態における検索の処理手順を示している。文書データ101は、レコード区切り文字で区切られた複数のレコード（文書）から成り、各レコードは、フィールド区切り文字で区切られた複数のフィールドから成っている。図３は文書データの具体例であり、フィールド区切り文字が「＾Ｆ」、レコード区切り文字が「＾Ｒ」で、紙名コード、日付、記事本文という３つのフィールドから成る新聞記事データである。

【００３１】単語頻度情報抽出手段103は、予め文書データ101を走査し、辞書102に登録されている単語が各レコードの記事本文フィールドに何回出現しているかをカウントし、当該単語が出現しているレコード数及び総レコード数とともに、単語頻度索引1304に格納する。

【００３２】また、書誌事項抽出手段105は、予め前記文書データ101を走査し、各レコードの書誌事項フィールドの内容を書誌事項索引106に格納する。

【００３３】まず、ステップ201：ユーザは検索要求入力手段107により、検索要求文字列を入力する。検索要求文字列は検索条件、ソート条件、適合度範囲指定の３つの部分からなる。図４は検索要求文字列の具体例を示しており、「松下 ＡＮＤ 新製品」の部分は検索条件で、「松下」と「新製品」という２つの検索語をともに記事本文に含むような記事を検索することを意味し、「@ＨＩＤＵＫＥ @ＳＨＩＭＥＩ」の部分はソート条件で、検索結果を日付の新しい順で並べ、同じ日付なら紙名コードの小さい順で並べるということを意味し、「＄７０：」の部分は適合度範囲指定で、適合度が最大である記事に対する相対適合度が７０以上である記事だけを結果に含めることを意味している。日付、紙名コードがどちらも同じ場合は適合度の順に並べる。なお、「＄７０：９０」のように適合度範囲指定の下限と上限とを両方指定して、適合度が７０以上９０以下の記事を結果に含めるといった指定や、上限だけを指定することも可能である。

【００３４】ステップ202：単語頻度算定手段108は、全ての検索語を対象として、

ステップ203：単語頻度索引104を参照し、検索要求入力手段107によって入力された検索条件に含まれる検索語について、当該単語が記事本文に出現するレコード数と各レコードの内部番号、各レコードにおける当該単語の出現頻度、及び総レコード数を算出する。

【００３５】ステップ204：論理演算手段109は、単語頻度算定手段108の出力したレコード集合間の論理演算を行う。図５は図４に示した検索要求文字列の場合の論理演算手段109の出力内容例を示しており、「松下」と「新製品」がともに出現するレコード集合が求められている。

【００３６】ステップ205：適合度算定手段110は、全ての検索結果レコードを対象として、

ステップ206：論理演算手段109の出力した各レコードについて、検索条件との適合度を、例えば、前記（数１）によって算出する。

【００３７】ステップ207：相対適合度算定手段111は、適合度算定手段110の出力した各レコードの適合度を、それらの最大値で除して１００倍した値に変換する。

【００３８】ステップ208：ソート情報取得手段112は、検索要求入力手段107で入力されたソート条件にしたがって書誌事項索引106を参照し、相対適合度算定手段111の出力した各レコードの、書誌事項の値をソート情報として取得する。図６はソート情報取得手段112の出力内容例で、日付と紙面コードの値をソート情報として取得している。

【００３９】ステップ209：中間結果足切り手段113は、ソート情報取得手段112から出力される全てのレコードを対象にして、

ステップ210：そのレコードの適合度が検索要求入力手段107から入力された適合度範囲指定に該当しているかをチェックし、

ステップ211：該当していないレコードは、除外する。

【００４０】図７は、適合度範囲指定が７０以上の場合に中間結果足切り手段113から出力される内容の例である。

【００４１】ステップ212：結果並べ替え手段114は、ソート情報として取得した複数の書誌事項をソートキーにして、中間結果足切り手段113の出力を並べ替え、全ての書誌事項の値が同じレコードの場合には適合度の大きい順に並べ替えて出力する。図８は、この結果並べ替え手段114の出力内容の例である。日付が新しく、紙名コードの小さい順に結果文書が並べられ、かつ、適合度が指定した範囲外だった記事は除外されているため、ユーザは効率良く所望の文書を見つけることができる。

【００４２】ステップ213：結果出力手段115は、結果並べ替え手段114の出力を整形してユーザに提示する。

【００４３】このように、この文書検索装置では、検索した文書の中から適合度範囲に入らない文書を除いて表示することができるため、所望の文書を効率よく探し出すことができる。

【００４４】また、検索結果の文書を適合度で足切りする場合に、検索結果を一旦適合度でソートし、適合度が所定値に満たない文書を足切りする方法も考えられるが、足切り前の検索結果の文書数は多いため、この文書を対象とするソートの処理負担は極めて重くなる。これ

に対して、この実施形態の方法では、文書の適合度が、指定された適合度範囲に入るかかどうかのチェックを、各文書に対して行うだけであるから、前記ソート処理に比べて軽い処理になる。従って、文書検索結果を迅速に表示することができる。

【００４５】なお、ステップ208のソート情報の取得は、ステップ209のＹＥＳの後、即ち、検索結果の足切りをした後の文書を対象に行うようにしても良く、そうした場合には、ソート情報の取得の作業量を減らすことができる。

【００４６】（第２の実施の形態）第２の実施形態では、適合度のランクで区別して文書を表示する文書検索装置について説明する。

【００４７】この装置は、図９に示すように、ソート情報取得手段912から出力された検索結果のレコードに対して、適合度の値が指定された適合度範囲に入るかどうかによって異なる区分けフラグを付与する中間結果区分け手段913を備えている。また、第１の実施形態と異なり、中間結果足切り手段は持たない。その他の構成は、第１の実施形態（図１）と変わりがない。

【００４８】図１０は、第２の実施形態における、検索の処理手順を示すフローチャートである。ここで、ステップ1008までの手順は、第１の実施形態と同様の処理手順である。

【００４９】ステップ1009：中間結果区分け手段913は、ソート情報取得手段912から出力される全てのレコードを対象にして、

ステップ1010：そのレコードの適合度が検索要求入力手段907から入力された適合度範囲指定に該当しているかをチェックし、

ステップ1011：適合度範囲に該当しないレコードについては区分けフラグの値として「２」を付与し、

ステップ1012：適合度範囲に該当するレコードについては区分けフラグの値として「１」を付与する。

【００５０】図１１は、中間結果区分け手段913の出力内容の例である。

【００５１】なお、適合度範囲として下限と上限の両方が指定された場合には、中間結果区分け手段913が、適合度範囲に該当しないレコードをさらに細分化して、上限を超えるレコードには区分けフラグの値として「２」を、下限に満たないレコードには区分けフラグの値として「３」を与えるようにしても良い。

【００５２】ステップ1013：結果並べ替え手段914は、中間結果区分け手段913の出力を、区分けフラグの値の降順で並べ替え、区分けフラグの値が同じだった場合には、ソート情報として取得した複数の書誌事項をソートキーとして並べ替え、すべての書誌事項の値が同じレコードがあった場合には適合度の大きい順に並べ替えて出力する。

【００５３】図１２は、結果並べ替え手段914の出力内容の例である。日付が新しく、紙名コードの小さい順に結果文書が並べられ、かつ、適合度が指定した範囲外だった記事は、適合度が指定範囲内にある記事群よりも下位にランクされるため、ユーザは効率良く所望の文書を見つけることができる。

【００５４】ステップ1014：結果出力手段915は、結果並べ替え手段914の出力を整形してユーザに提示する。

【００５５】このように、この実施形態の文書検索装置では、検索された全ての文書を、適合度範囲に入るものと入らないものとに区分して表示することができる。ユーザは、検索の目的に応じて、適合度範囲に該当する区分の文書だけを見て文書検索を終了することもできるし、特許文書を検索するときのように、１つの漏れも許されない場合には、適合度範囲から外れる区分の文書についても逐一調べることが可能である。

【００５６】

【発明の効果】以上の説明から明らかなように、本発明の文書検索装置及び文書検索方法では、適合度がユーザの指定した範囲から外れる文書を、検索結果から除いたり、より下位にランク付けすることができる。

【００５７】そうすることにより、ソート条件を指定した場合の、適合度の低い文書が上位に、適合度の高い文書が下位にランクされてしまうという問題を回避でき、所望の文書を効率良く検索することが可能になる。

【００５８】また、各文書の適合度を最大値に対する相対値に変換し、検索要求における適合度範囲指定も相対値で指定することにより、適切な適合度範囲を容易に指定できる。

【図面の簡単な説明】

【図１】本発明の第１の実施の形態における文書検索装置の構成を示すブロック図、

【図２】第１の実施の形態における検索処理の手順を示す流れ図、

【図３】文書データの一例を示す図、

【図４】第１の実施形態における検索要求文字列の一例を示す図、

【図５】第１の実施形態における論理演算手段の出力内容の一例を示す図、

【図６】第１の実施形態におけるソート情報取得手段の出力内容の一例を示す図、

【図７】第１の実施形態における中間結果足切り手段の出力内容の一例を示す図、

【図８】第１の実施形態における結果並べ替え手段の出力内容の一例を示す図、

【図９】本発明の第２の実施の形態における文書検索装置の構成を示すブロック図、

【図１０】第２の実施の形態における検索処理の手順を示す流れ図、

【図１１】第２の実施形態における中間結果区分け手段の出力内容の一例を示す図、

11

【図１２】第２の実施形態における結果並べ替え手段の出力内容の一例を示す図、

【図１３】従来の文書検索装置の構成を示すブロック図、

【図１４】従来の検索処理の手順を示す流れ図、

【図１５】検索要求文字列の一例を示す図、

【図１６】結果並べ替え手段の出力内容の一例を示す図である。

【符号の説明】

101、901、1301　文書データ

102、902、1302　辞書

103、903、1303　単語頻度情報抽出手段

104、904、1304　単語頻度索引

12

105、905、1305　書誌事項抽出手段

106、906、1306　書誌事項索引

107、907、1307　検索要求入力手段

108、908、1308　単語頻度算定手段

109、909、1309　論理演算手段

110、910、1310　適合度算定手段

111、911　相対適合度算定手段

112、912、1311　ソート情報取得手段

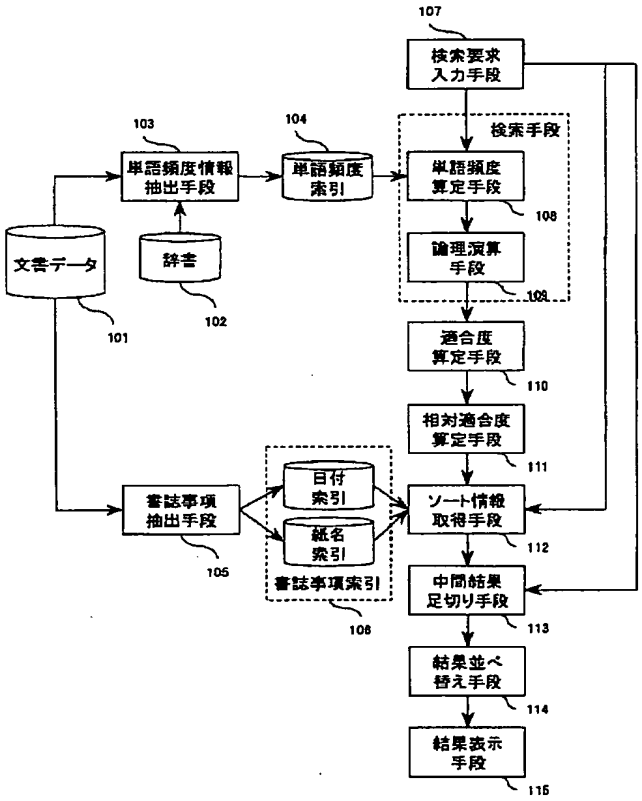
113　中間結果足切り手段

114、914、1312　結果並べ替え手段

115、915、1313　結果表示手段

913　中間結果区分け手段

【図１】



【図３】

05^F19870317^FNY円急反落(海外外為)。十六日のニューヨーク外国為替市場の円相場は急反落。・・・^R02^F19870317^F少年のナイフ事件続発、・・・

【図５】

レコード内部番号	「松下」のレコード内出現頻度	「新製品」のレコード内出現頻度
10	1	2
80	3	3
・	・	・
・	・	・
・	・	・
・	・	・
2,100,255	4	2
2,100,256	2	4
2,100,257	4	1
2,100,259	6	5
2,100,361	2	1
2,100,365	1	2

3289件

総レコード数	2,100,865
「松下」の出現レコード数	13,271
「新製品」の出現レコード数	8,280

【図４】

松下 AND 新製品 @HIDUKE @SHIMEI $70:

【図１５】

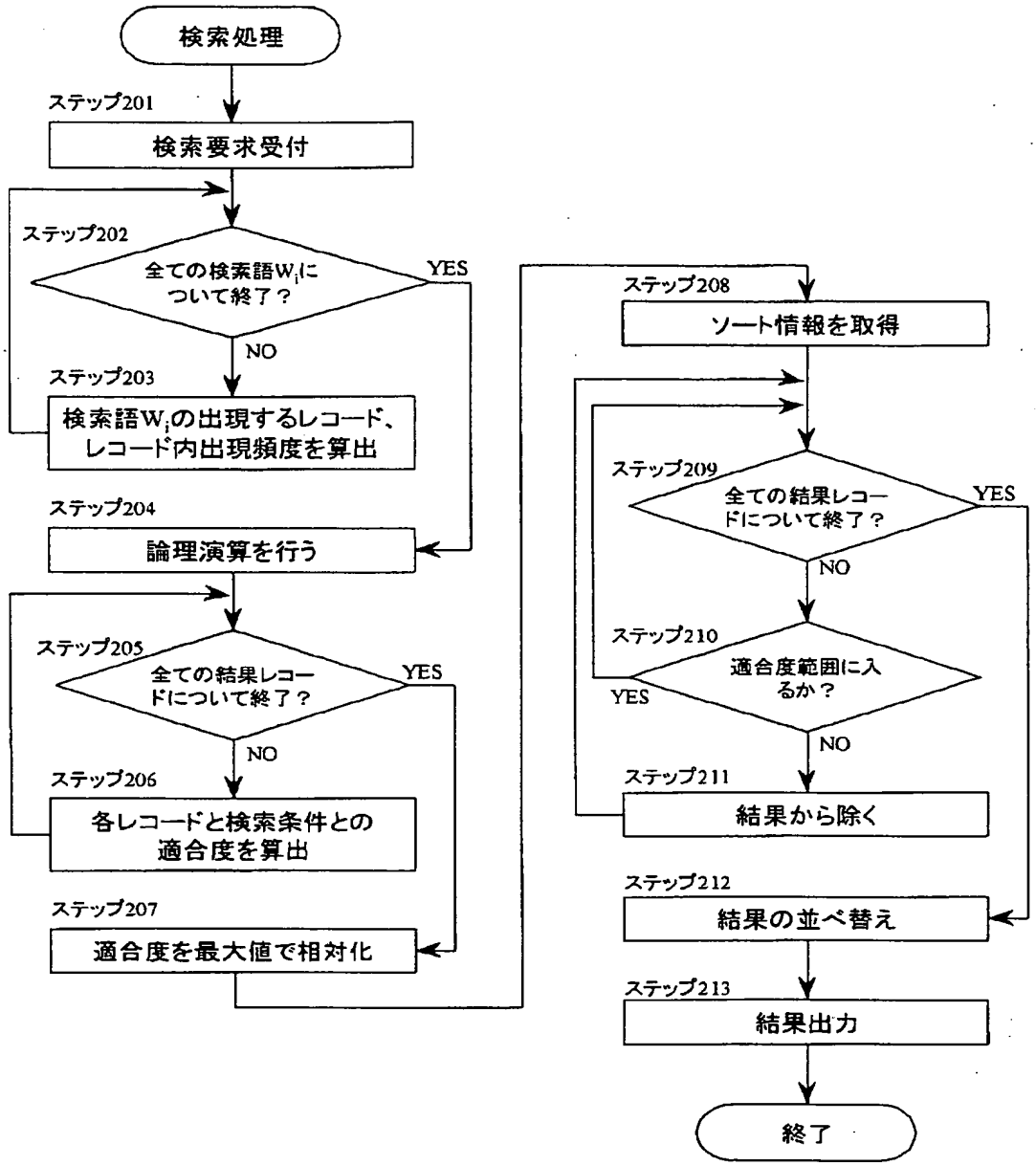
松下 AND 新製品 @HIDUKE @SHIMEI

【図６】

レコード内部番号	日付	紙名コード	適合度
10	19870502	02	84.1
80	19870502	01	50.2
・	・	・	・
・	・	・	・
・	・	・	・
・	・	・	・
2,100,255	19990725	05	77.5
2,100,256	19990725	05	70.6
2,100,257	19990725	05	100
2,100,259	19990725	05	1.5
2,100,361	19990725	01	88.8
2,100,365	19990725	01	5.3

3289件

【図２】



【図７】

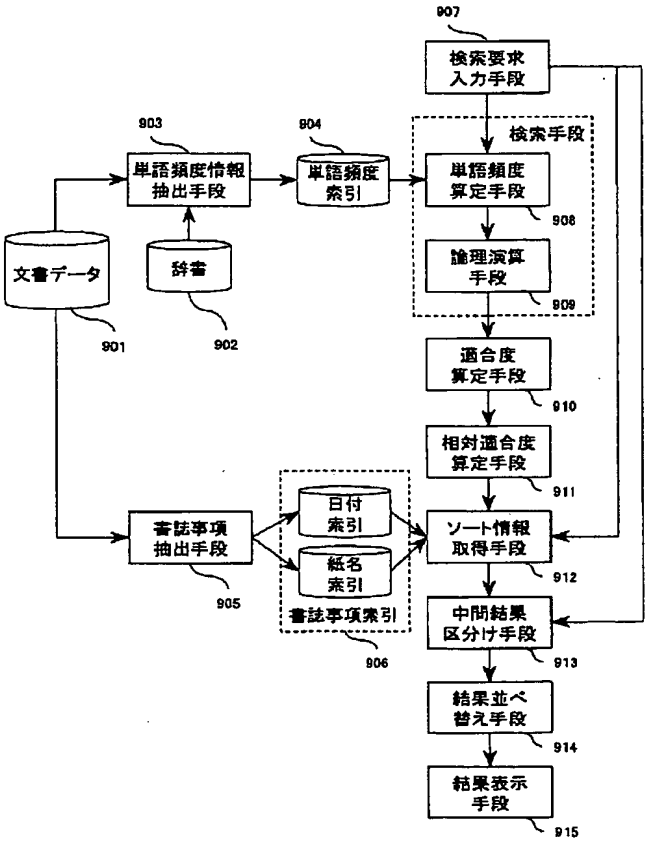
レコード内部番号	日付	紙名コード	適合度
10	19870502	02	84.1
・	・	・	・
・	・	・	・
2,100,255	19990725	05	77.5
2,100,256	19990725	05	70.6
2,100,257	19990725	05	100
2,100,381	19990725	01	88.8

82件

【図８】

順位	レコード内部番号	日付	紙名コード	適合度
1	2,100,381	19990725	01	88.8
2	2,100,257	19990725	05	100
3	2,100,255	19990725	05	77.5
4	2,100,256	19990725	05	70.6
・	・	・	・	・
・	・	・	・	・
82	10	19870502	02	84.1

【図９】



【図１１】

レコード内部番号	日付	紙名コード	適合度	区分けフラグ
10	19870502	02	84.1	1
80	19870502	01	50.2	2
・	・	・	・	
・	・	・	・	
・	・	・	・	
・	・	・	・	
2,100,255	19990725	05	77.5	1
2,100,256	19990725	05	70.6	1
2,100,257	19990725	05	100	1
2,100,259	19990725	05	1.5	2
2,100,361	19990725	01	88.8	1
2,100,365	19990725	01	5.3	2

3289件

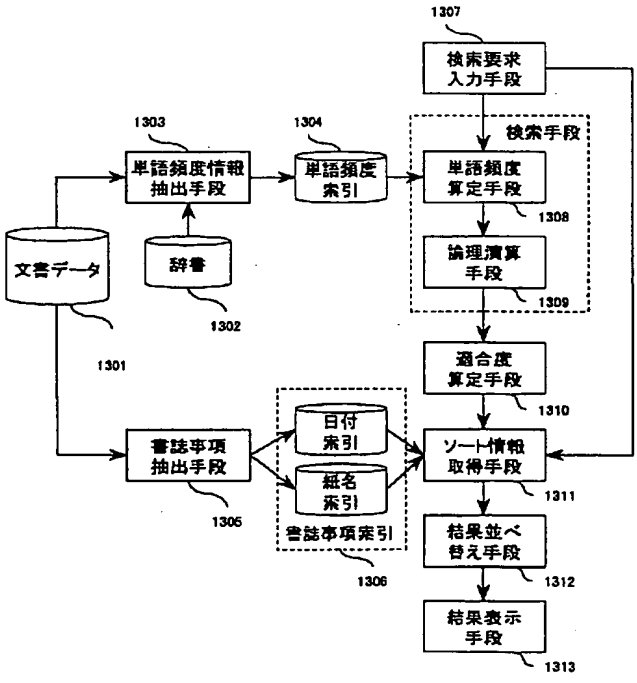
【図１６】

順位	レコード内部番号	日付	紙名コード	適合度
1	2,100,361	19990725	01	8.88
2	2,100,365	19990725	01	0.53
3	2,100,257	19990725	05	10
4	2,100,255	19990725	05	7.75
5	2,100,256	19990725	05	7.06
6	2,100,259	19990725	05	0.15
・	・	・	・	・
・	・	・	・	・
・	・	・	・	・
・	・	・	・	・
3288	80	19870502	01	5.02
3289	10	19870502	02	8.41

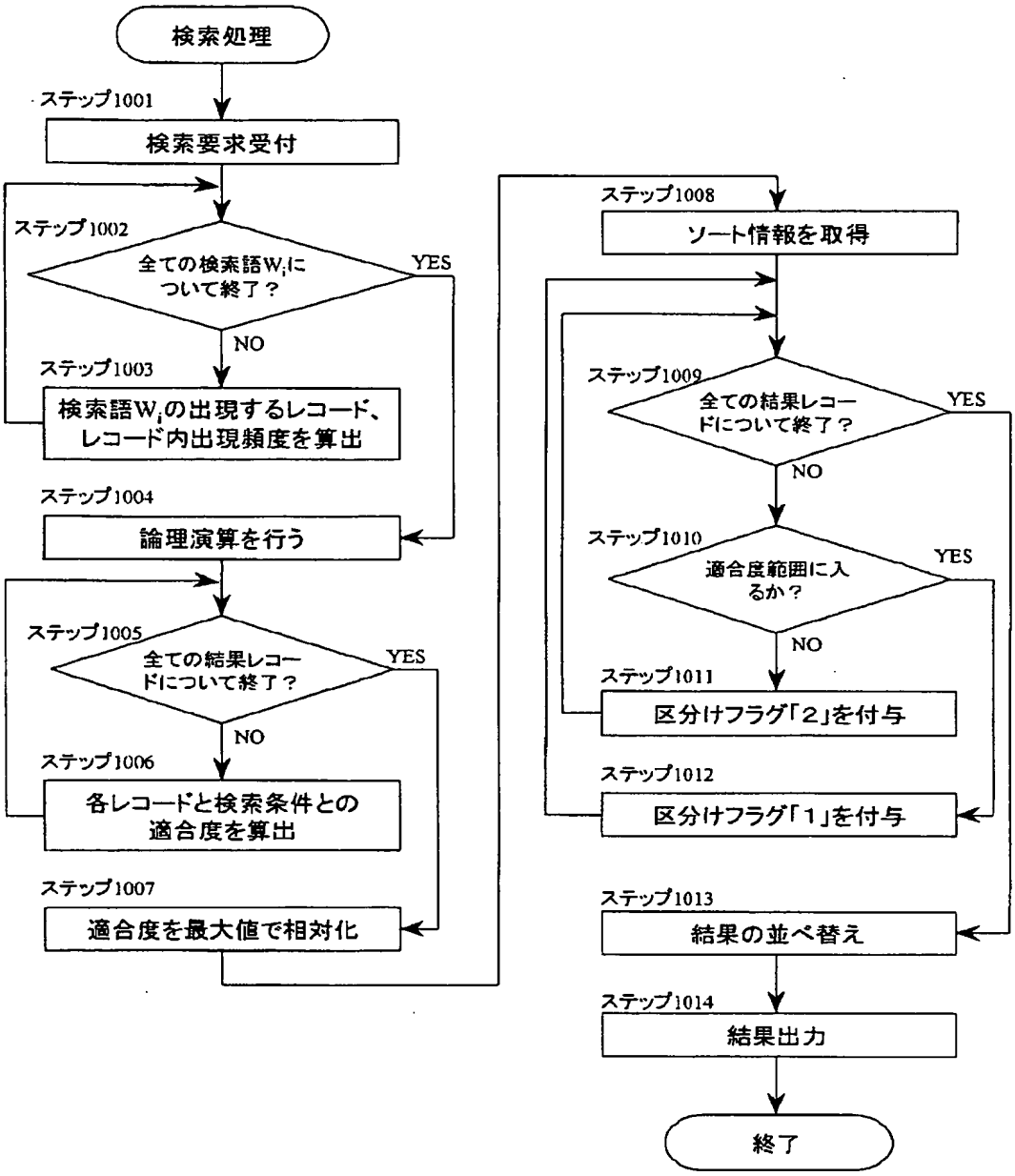
【図１２】

順位	レコード内部番号	日付	紙名コード	適合度	区分けフラグ
1	2,100,361	19990725	01	88.8	1
2	2,100,257	19990725	05	100	1
3	2,100,255	19990725	05	77.5	1
4	2,100,256	19990725	05	70.6	1
・	・	・	・	・	・
・	・	・	・	・	・
82	10	19870502	02	84.1	1
83	2,100,365	19990725	01	5.3	2
84	2,100,259	19990726	05	1.5	2
・	・	・	・	・	・
・	・	・	・	・	・
・	・	・	・	・	・
3289	80	19870502	01	50.2	2

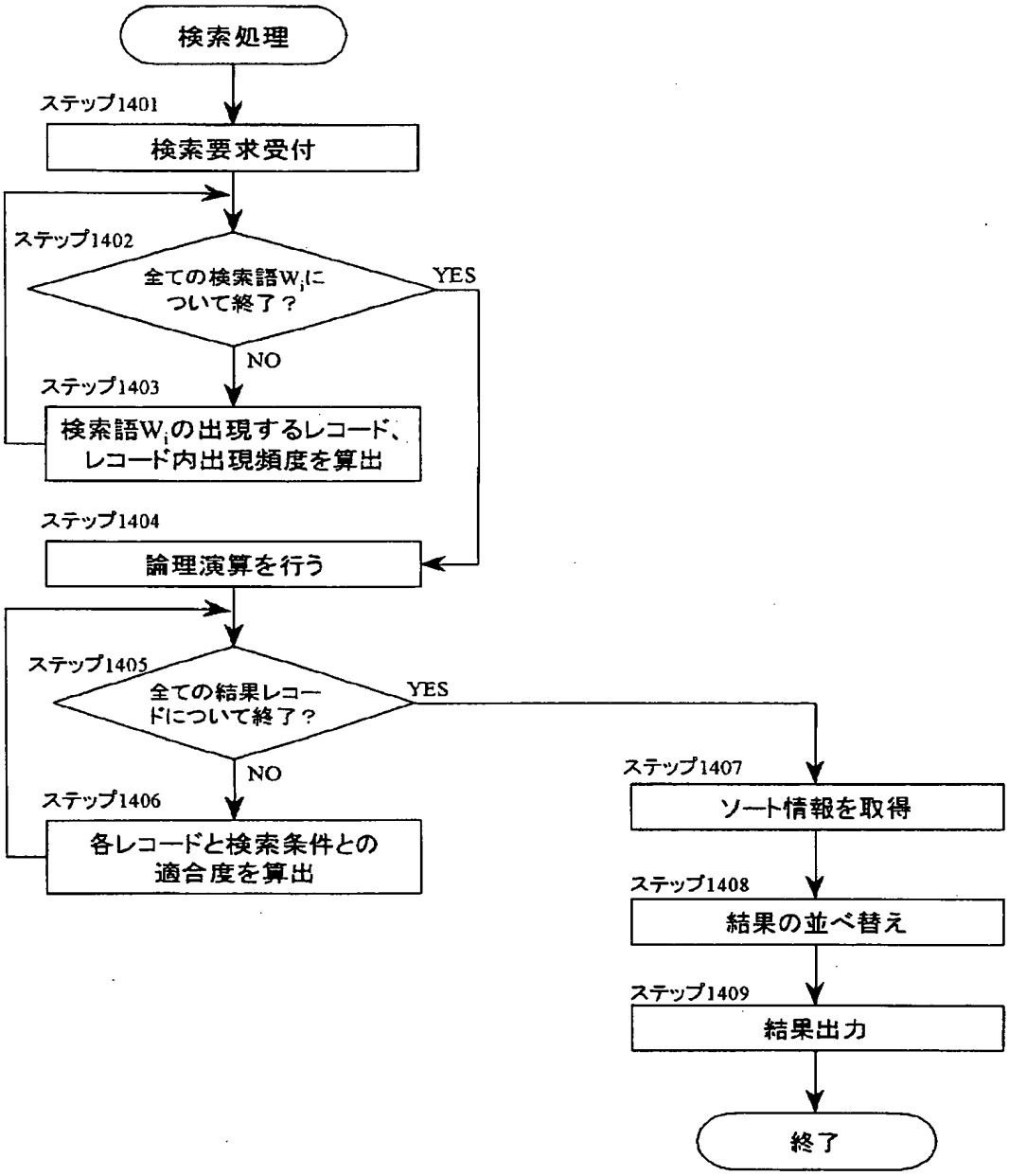
【図１３】



【図１０】


検索処理
ステップ1001
検索要求受付
ステップ1002
全ての検索語Wiについて終了？
YES
NO
ステップ1003
検索語Wiの出現するレコード、レコード内出現頻度を算出
ステップ1004
論理演算を行う
ステップ1005
全ての結果レコードについて終了？
YES
NO
ステップ1006
各レコードと検索条件との適合度を算出
ステップ1007
適合度を最大値で相対化
ステップ1008
ソート情報を取得
ステップ1009
全ての結果レコードについて終了？
YES
NO
ステップ1010
適合度範囲に入るか？
YES
NO
ステップ1011
区分けフラグ「2」を付与
ステップ1012
区分けフラグ「1」を付与
ステップ1013
結果の並べ替え
ステップ1014
結果出力
終了

【図１４】


検索処理
ステップ1401
検索要求受付
ステップ1402
全ての検索語W_iについて終了？
YES
NO
ステップ1403
検索語W_iの出現するレコード、レコード内出現頻度を算出
ステップ1404
論理演算を行う
ステップ1405
全ての結果レコードについて終了？
YES
NO
ステップ1406
各レコードと検索条件との適合度を算出
ステップ1407
ソート情報を取得
ステップ1408
結果の並べ替え
ステップ1409
結果出力
終了

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

CLAIMS

[Claim(s)]
[Claim 1]A document retrieval system which searches accumulated document data according to a search condition, and rearranges and displays search results according to a sort condition, comprising:
A search condition.
A sort condition.
A retrieval-required input means which inputs a retrieval-required character string which comprises range specification of goodness of fit which shows a degree corresponding to said search condition.
A search means to search a document which fulfills said search condition.
A goodness-of-fit calculating means which computes said goodness of fit of each document searched by said search means.
A sorting information acquisition means which acquires sorting information for rearranging about said each searched document according to said sort condition.
A search-results cut-off point means except a document in which goodness of fit does not go into said goodness-of-fit range by which range specification was carried out from said each searched document, A search-results rearrangement means which rearranges by said sorting information, and first rearranges each document included in said goodness-of-fit range outputted from said search-results cut-off point means in order of said goodness of fit when said sorting information is the same, A search-results displaying means which displays search results rearranged by said search-results rearrangement means.

[Claim 2]A document retrieval system which searches accumulated document data according to a search condition, and rearranges and displays search results according to a sort condition, comprising:

A search condition.
A sort condition.
A retrieval-required input means which inputs a retrieval-required character string which comprises range specification of goodness of fit which shows a degree corresponding to said search condition.
A search means to search a document which fulfills said search condition.
A goodness-of-fit calculating means which computes said goodness of fit of each document searched by said search means.
A sorting information acquisition means which acquires sorting information for rearranging about said each searched document according to said sort condition.
A search-results division means to give a division flag which shows whether goodness of fit of each of said searched document is gone into said goodness-of-fit range as compared with said goodness-of-fit range by which range specification was carried out to said each document,
First, rearrange with said division flag, and when a value of said division flag is the same, each document in which said division flag outputted from said search-results division means was given, A search-results rearrangement means which is rearranged by said sorting information, and is rearranged in order of said goodness of fit when said sorting information is the same, and a search-results displaying means which displays search results rearranged by said search-results rearrangement means.

[Claim 3]While searching a document corresponding to said search condition, said search means computes the frequency of occurrence of a search term in each document, and said goodness-of-fit calculating means, The document retrieval system according to claim 1 or 2 computing said goodness of fit of each document based on the frequency of occurrence of a search term computed by said search means.
[Claim 4]While searching a document corresponding to said search condition, said search means computes a document number in which a search term appears, and the frequency of occurrence of a search term in each document, and said goodness-of-fit calculating means, The document retrieval system according to claim 1 or 2 computing said goodness of fit of each document based on the frequency of occurrence of a search term in each document, and an appearance document number of a search term.
[Claim 5]The document retrieval system according to claim 1 or 2 which it has the following, and said goodness-of-fit calculating means outputs relative goodness of fit expressed with said relative value as goodness of fit of each document, and is characterized by said retrieval-required input means performing said range specification of goodness of fit with said relative goodness of fit.
An absolute goodness-of-fit calculation means as which said goodness-of-fit calculating means

calculates goodness of fit of each document.
A relative goodness-of-fit calculation means which changes into a relative value over the highest goodness of fit of them goodness of fit of each document calculated by said absolute goodness-of-fit calculation means.

[Claim 6]In a document retrieval method which searches accumulated document data according to a search condition, and rearranges and displays search results according to a sort condition, As opposed to retrieval required which specifies a search condition, a sort condition, and the range of goodness of fit which shows a degree corresponding to said search condition, Search a document which fulfills said search condition from accumulated document data, and said goodness of fit of each detected document is computed, Sorting information of each of said document for rearranging said each document according to said sort condition is acquired, A document retrieval method rearranging by said sorting information, first rearranging each document which remained in search results except for a document which does not go into the range of goodness of fit specified by said retrieval required from search results in order of said goodness of fit when said sorting information is the same, and displaying it.
[Claim 7]In a document retrieval method which searches accumulated document data according to a search condition, and rearranges and displays search results according to a sort condition, As opposed to retrieval required which specifies a search condition, a sort condition, and the range of goodness of fit which shows a degree corresponding to said search condition, Search a document which fulfills said search condition from accumulated document data, and said goodness of fit of each detected document is computed, A document which does not go into the range of goodness of fit specified by said retrieval required is removed from search results, A document retrieval method characterized by what sorting information of each of said document for rearranging each document which remained in search results according to said sort condition is acquired, and said each document is first rearranged by said sorting information, it rearranges in order of said goodness of fit when said sorting information is the same, and is displayed.
[Claim 8]In a document retrieval method which searches accumulated document data according to a search condition, and rearranges and displays search results according to a sort condition, As opposed to retrieval required which specifies a search condition, a sort condition, and the range of goodness of fit which shows a degree corresponding to said search condition, Search a document which fulfills said search condition from accumulated document data, and said goodness of fit of each detected document is computed, Sorting information of each of said document for rearranging said each document according to said sort condition is acquired, A division flag which shows whether goodness of fit of each of said document is

gone into said range as compared with the range of goodness of fit specified by said retrieval required is given to said each document, A document retrieval method it rearranges with said division flag, rearranging by said sorting information first when a value of said division flag is the same, rearranging said each document in order of said goodness of fit when said sorting information is the same, and displaying it.

[Claim 9]Claim 6 characterized by computing relative goodness of fit to the highest goodness of fit of the goodness of fit of each of said document, and enabling it to specify the range of goodness of fit with said relative goodness of fit in said retrieval required as said goodness of fit of each detected document, the document retrieval method according to claim 7 or 8.

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[Field of the Invention]The degree with which each document agrees in a search condition especially about the document retrieval system and document retrieval method with which this invention searches a desired document according to a search condition, If it is the bibliographic items which accompany each document, for example, a newspaper article, search results will be rearranged and it will enable it to display with combination, such as new order of the date.
[0002]
[Description of the Prior Art]In recent years, based on the frequency of occurrence of the search term in a document, etc., the technique of the document ranking which asks for the goodness of fit of a document and a search condition, and rearranges and displays a result on the high order has attracted attention. If it is the bibliographic items which accompany a document, for example, a newspaper article, the date will be specified as a sort condition, and it will give priority and display from the new report of the date, but about the report of the same date, flexible search [say / displaying on order with high goodness of fit with a search condition] has been realized.
[0003]The word frequency information extraction means 1303 which extracts the word frequency information on the word which appears in the dictionary 1302 from the document data 1301 of the newspaper article used as a retrieval object, and is stored in the word frequency index 1304 as the conventional document retrieval system is shown in drawing 13, A bibliographic-items extraction means 1305 to take out the information on bibliographic items, such as a date and a paper name code, from the document data 1301, and to store in the bibliographic-items index 1306, The retrieval-required input means 1307 for a user to input the retrieval-required character string which consists of a search condition and a sort condition, The word frequency calculation means 1308 which asks for the frequency of occurrence in the

inside of the document of the search term which investigates the word frequency index 1304 and is contained in a search condition, The logic operation means 1309 which performs the logical operation during a record set, and the goodness-of-fit calculation means 1310 which computes the goodness of fit of a search condition and each record, As a result of rearranging the record list of search results with the sorting information acquisition means 1311 which acquires the bibliographic information for rearrangement specified as the sort condition, and bibliographic information and goodness of fit, it has the rearrangement means 1312 and a result display means 1313 to display search results.

[0004]The frequency of occurrence of a dictionary word retrieval object sentence in the letter is stored in the word frequency index 1304 by the extraction operation of the word frequency information extraction means 1303.

[0005]Drawing 14 is a flow chart which shows the procedure of the search in the conventional document retrieval system. The document data 1301 comprises two or more records (document) divided with the record delimiter, and each record comprises two or more fields divided with the field delimiter. Drawing 3 shows the example of document data, and a field delimiter is "^F", and a record delimiter is "^R", and it is newspaper article data which comprises the three fields called a paper name code, a date, and the body of story.

[0006]The word frequency information extraction means 1303 scans the document data 1301 beforehand, It counts how many times the word registered into the dictionary 1302 has appeared in the body-of-story field of each record, and stores in the word frequency index 1304 with the record count in which the word concerned has appeared, and the total record count.

[0007]The bibliographic-items extraction means 1305 scans the document data 1301 beforehand, and stores the contents of the bibliographic-items field of each record in the bibliographic-items index 1306.

[0008]First, the step 1401: A user inputs a retrieval-required character string by the retrieval-required input means 1307. A retrieval-required character string consists of two portions, a search condition and a sort condition. Drawing 15 shows the example of the retrieval-required character string, and is "Matsushita AND. The portion of new product" is a search condition, It means searching a report which contains in the body of story both the two search terms "Matsushita" and a "new product", and the portion of "@HIDUKE @SHIMEI" is a sort condition and it means putting search results in order in the new order of the date, and putting the same date in order in order with a small paper name code. Both of dates and paper name codes arrange in order of goodness of fit, when the same.

[0009]Step 1402 : the word frequency calculation means 1308, The step 1403:word frequency index 1304 is referred to for all the search terms, About the search term contained in the search condition inputted by the retrieval-required input means 1307, the word concerned

computes the frequency of occurrence and the total record count of the internal number of the record count which appears in the body of story, and each record, and the word concerned in each record.

[0010]Step 1404: The logic operation means 1309 performs the logical operation during the record set which the word frequency calculation means 1308 outputted.

[0011]Step 1405: The goodness-of-fit calculation means 1310 computes goodness of fit (Rel) with a search condition, for example (several 1) about each record which the step 1406:logic operation means 1309 outputted for all the search-results records.

Rel = sigma(TFi-IDFi) (sigma adds about i)

IDFi =1-$\log_2$(DFi/ND) (several 1)

However, the record count and ND in which, as for TFi, the frequency of occurrence in a record of the search term Wi appears, and, as for DFi, the word Wi appears express the total record count.

[0012]The calculating method of goodness of fit is not restricted to (several 1).

[0013]Step 1407: The sorting information acquisition means 1311 acquires the value of the bibliographic items corresponding to the sort condition inputted from the retrieval-required input means 1307 of each record which the goodness-of-fit calculation means 1310 outputted as sorting information with reference to the bibliographic-items index 1306.

[0014]Drawing 6 shows the example of the contents of an output of the sorting information acquisition means 1311, and acquires the value of the date and a space code as sorting information.

[0015]Step 1408: In a result, the rearrangement means 1312 rearranges and outputs the output of the sorting information acquisition means 1311 by making into a sort key two or more bibliographic items acquired as sorting information. At this time, when there is the record with same value of all the bibliographic items, it rearranges into descending of goodness of fit.

[0016]There is drawing 16 in the example of the contents of an output of the result rearrangement means 1312.

[0017]Step 1409: The result display means 1313 operates the output of the result rearrangement means 1312 orthopedically, and shows it to a user.

[0018]

[Problem(s) to be Solved by the Invention]However, since priority is given over goodness of fit to the value etc. of the bibliographic items specified as the sort condition as a key of rearrangement in the conventional composition, A document with low goodness of fit will be ranked as a higher rank, a document with high goodness of fit may be ranked as a low rank, and there was a problem that a desired document was efficiently nondiscoverable.

[0019]For example, the document (record internal number 10) ranked as the lowest in drawing 8 hits this.

[0020]Though seriously taken as a key of rearrangement of the value of the bibliographic items which this invention solves the technical problem of such conventional technology, and were specified as the sort condition, It aims at providing the document retrieval system which can discover a desired document efficiently, and providing the document retrieval method by removing from a result the document which does not go into the goodness-of-fit range which the user could limit the range of goodness of fit, and specified, or ranking it as a low rank more.

[0021]
[Means for Solving the Problem]So, in addition to a search condition and a sort condition, in a document retrieval system of this invention, a retrieval-required input means which inputs goodness-of-fit range specification, and a search-results cut-off point means excluding a document in which goodness of fit does not go into a specified goodness-of-fit range from search results are formed as a retrieval-required character string.
[0022]A retrieval-required input means which inputs goodness-of-fit range specification as a retrieval-required character string in addition to a search condition and a sort condition, and goodness of fit of a document have formed an intermediate result division means to give a different division flag, by whether it corresponds to a specified goodness-of-fit range.
[0023]In a document retrieval method of this invention, retrieval required which specifies a search condition, a sort condition, and the range of goodness of fit is received, Search a document which fulfills a search condition from accumulated document data, and goodness of fit of each searched document is computed, It rearranges by sorting information, and when sorting information is the same, he rearranges each document which remained in search results except for a document which does not go into the range of goodness of fit specified by retrieval required from search results in order of goodness of fit, and is trying to acquire sorting information of each document and to display it first.
[0024]A procedure which acquires sorting information of each of this document, and a procedure excluding a document which does not go into the range of goodness of fit specified by retrieval required from search results are replaced.
[0025]Retrieval required which specifies a search condition, a sort condition, and the range of goodness of fit which shows a degree corresponding to a search condition is received, Search a document which fulfills a search condition from accumulated document data, and goodness of fit of each detected document is computed, Acquire sorting information of each document for rearranging each document according to a sort condition, and goodness of fit of each document is measured with the range of goodness of fit specified by retrieval required, A division flag which shows whether it goes into the range is given to each document, by rearranging with a division flag first, and rearranging by sorting information, when a value of a division flag is the same, each document is rearranged in order of goodness of fit, when sorting

information is the same, and he is trying to display.
[0026]Therefore, goodness of fit can remove a document which separates from a range which a user specified from search results, or it can be ranked more as a low rank, and a problem that a document with low goodness of fit at the time of specifying a sort condition will be ranked as a higher rank, and a document with high goodness of fit will be ranked as a low rank can be avoided.
[0027]
[Embodiment of the Invention]Hereafter, an embodiment of the invention is described, referring to a figure.
[0028](A 1st embodiment) Drawing 1 is a block diagram showing the composition of the document retrieval system in a 1st embodiment of this invention.
[0029]This device is provided with the following.
Like the conventional device (drawing 13), The word frequency information on the word which appears in the dictionary 102 is extracted from the document data 101 of the newspaper article used as a retrieval object. The information on bibliographic items, such as a date and a paper name code, is taken out from the word frequency information extraction means 103 stored in the word frequency index 104, and the document data 101. a bibliographic-items extraction means 105 to store in the bibliographic-items index 106, the retrieval-required input means 107, the word frequency calculation means 108, the logic operation means 109, the goodness-of-fit calculation means 110, the sorting information acquisition means 112, and a result -- the rearrangement means 114 -- and, The relative goodness-of-fit calculation means 111 which changes into the relative value over the maximum the goodness of fit of each record calculated by the goodness-of-fit calculation means 110, and is outputted to the sorting information acquisition means 112 while having the result display means 115.
The intermediate result cut-off point means 113 except the record which does not go into the goodness-of-fit range which the value of goodness of fit specified from the search results outputted from the sorting information acquisition means 112.

[0030]The flow chart of drawing 2 shows the procedure of the search in a 1st embodiment. The document data 101 comprises two or more records (document) divided with the record delimiter, and each record comprises two or more fields divided with the field delimiter. It is an example of document data, and a field delimiter is "^F", a record delimiter is "^R", and drawing 3 is newspaper article data which comprises the three fields called a paper name code, a date, and the body of story.
[0031]The word frequency information extraction means 103 scans the document data 101 beforehand, counts how many times the word registered into the dictionary 102 has appeared in the body-of-story field of each record, and stores it in the word frequency index 1304 with

the record count in which the word concerned has appeared, and the total record count.
[0032]The bibliographic-items extraction means 105 scans said document data 101 beforehand, and stores the contents of the bibliographic-items field of each record in the bibliographic-items index 106.
[0033]First, the step 201: A user inputs a retrieval-required character string by the retrieval-required input means 107. A retrieval-required character string consists of three portions, a search condition, a sort condition, and goodness-of-fit range specification. Drawing 4 shows the example of the retrieval-required character string, and is "Matsushita AND. The portion of new product" is a search condition, It means searching a report which contains in the body of story both the two search terms "Matsushita" and a "new product", and the portion of "@HIDUKE @SHIMEI" is a sort condition, It means putting search results in order in the new order of the date, and putting the same date in order in order with a small paper name code, and means that the portion of "$70:" includes in a result only the report whose relative goodness of fit to the report whose goodness of fit it is goodness-of-fit range specification, and is the maximum is 70 or more. Both of dates and paper name codes arrange in order of goodness of fit, when the same. It is also possible to specify both of the minimums and maximums of goodness-of-fit range specification, as shown in "$70:90", and to specify only specification that goodness of fit includes or more 70 90 or less report in a result, and a maximum.
[0034]Step 202 : the word frequency calculation means 108, The step 203:word frequency index 104 is referred to for all the search terms, About the search term contained in the search condition inputted by the retrieval-required input means 107, the word concerned computes the frequency of occurrence and the total record count of the internal number of the record count which appears in the body of story, and each record, and the word concerned in each record.
[0035]Step 204: The logic operation means 109 performs the logical operation during the record set which the word frequency calculation means 108 outputted. Drawing 5 shows the example of the contents of an output of the logic operation means 109 in the case of the retrieval-required character string shown in drawing 4, and the record set in which both "Matsushita" and a "new product" appear is called for.
[0036]Step 205: The goodness-of-fit calculation means 110 computes goodness of fit with a search condition with the above (several 1), for example about each record which the step 206:logic operation means 109 outputted for all the search-results records.
[0037]Step 207: Change the relative goodness-of-fit calculation means 111 into the value which **(ed) goodness of fit of each record which the goodness-of-fit calculation means 110 outputted at those maximums, and increased it 100 times.
[0038]Step 208: The sorting information acquisition means 112 acquires the value of bibliographic items of each record which the relative goodness-of-fit calculation means 111

outputted with reference to the bibliographic-items index 106 according to the sort condition inputted by the retrieval-required input means 107 as sorting information. Drawing 6 is an example of the contents of an output of the sorting information acquisition means 112, and acquires the value of the date and a space code as sorting information.

[0039]Step 209 : the intermediate result cut-off point means 113, being aimed at all the records outputted from the sorting information acquisition means 112 -- step 210: -- the record which confirms whether the goodness of fit of the record corresponds to the goodness-of-fit range specification inputted from the retrieval-required input means 107, and has not done step 211:relevance of it is excepted.

[0040]Drawing 7 is an example of the contents outputted from the intermediate result cut-off point means 113, when goodness-of-fit range specification is 70 or more.

[0041]Step 212: The rearrangement means 114 makes a sort key two or more bibliographic items acquired as sorting information, and rearranges the output of the intermediate result cut-off point means 113, and a result rearranges and outputs it to descending of goodness of fit, when the value of all the bibliographic items is the same record. As a result, drawing 8 is an example of the contents of an output of the rearrangement means 114. order with the new date and a small paper name code -- a result -- a document -- arranging -- having -- and since the report which was outside the range specified by goodness of fit is excepted, the user can find a desired document efficiently.

[0042]Step 213: The result outputting means 115 operates the output of the result rearrangement means 114 orthopedically, and shows it to a user.

[0043]Thus, since it can display except for the document which does not go into the goodness-of-fit range out of the document searched with this document retrieval system, a desired document is efficiently discoverable.

[0044]When cutting off the document of search results with goodness of fit, search results are once sorted with goodness of fit, and the way goodness of fit cuts off the document which is less than a predetermined value is also considered, but since there are many document numbers of the search results before a cut-off point, the processing burden of sorting for this document becomes very heavy. or [on the other hand, / that the goodness of fit of a document goes into the specified goodness-of-fit range in the method of this embodiment] -- since ****** is only checked to each document, compared with said sorting application, it becomes light processing. Therefore, a document-retrieval result can be displayed promptly.

[0045]It may be made to perform acquisition of the sorting information of Step 208 for the document after YES of Step 209 (i.e., after carrying out the cut-off point of search results), and, in such a case, the rating of acquisition of sorting information can be reduced.

[0046](A 2nd embodiment) A 2nd embodiment explains the document retrieval system which distinguishes by the rank of goodness of fit and displays a document.

[0047]This device is provided with an intermediate result division means 913 to give a division flag which is different by whether it goes into the goodness-of-fit range as which the value of goodness of fit was specified to the record of the search results outputted from the sorting information acquisition means 912 as shown in drawing 9. Unlike a 1st embodiment, it does not have an intermediate result cut-off point means. Other composition does not have a 1st embodiment (drawing 1) and a change.
[0048]Drawing 10 is a flow chart in a 2nd embodiment which shows the procedure of search. Here, the procedure to Step 1008 is the same procedure as a 1st embodiment.
[0049]Step 1009 : the intermediate result division means 913, It is aimed at all the records outputted from the sorting information acquisition means 912, Step 1010 : It is confirmed whether the goodness of fit of the record corresponds to the goodness-of-fit range specification inputted from the retrieval-required input means 907, Step 1011: Give "2" as a value of a division flag about the record which does not correspond to the goodness-of-fit range, and give "1" as a value of a division flag about the record applicable to the step 1012:goodness-of-fit range.
[0050]Drawing 11 is an example of the contents of an output of the intermediate result division means 913.
[0051]When both a minimum and a maximum are specified as a goodness-of-fit range, The intermediate result division means 913 subdivides further the record which does not correspond to the goodness-of-fit range, and it may be made to give "3" to the record which is less than a minimum in "2" as a value of a division flag at the record exceeding a maximum as a value of a division flag.
[0052]As a result of Step 1013:, the rearrangement means 914, Rearrange in descending order of the value of a division flag, and when the value of a division flag is the same, the output of the intermediate result division means 913, Two or more bibliographic items acquired as sorting information are rearranged as a sort key, and when there is the record with same value of all the bibliographic items, it rearranges and outputs to descending of goodness of fit.
[0053]There is drawing 12 in the example of the contents of an output of the result rearrangement means 914. order with the new date and a small paper name code -- a result -- a document -- arranging -- having -- and since the report which was outside the range specified by goodness of fit is ranked as a low rank rather than article groups which has goodness of fit in a designated range, the user can find a desired document efficiently.
[0054]Step 1014: The result outputting means 915 operates the output of the result rearrangement means 914 orthopedically, and shows it to a user.
[0055]Thus, all the documents searched with the document retrieval system of this embodiment can be classified and displayed on the thing included in the goodness-of-fit range, and the thing which does not enter. The user can also end a document retrieval, being able to

see only the document of the classification applicable to the goodness-of-fit range according to the purpose of search, and like [when searching a patent document], when one leakage is not allowed, either, he can investigate in detail also about the divisional document which separates from the goodness-of-fit range.

[0056]

[Effect of the Invention]In the document retrieval system and document retrieval method of this invention, goodness of fit can remove the document which separates from the range which the user specified from search results, or it can be ranked more as a low rank so that clearly from the above explanation.

[0057]By doing so, the problem that the document with low goodness of fit at the time of specifying a sort condition will be ranked as a higher rank, and a document with high goodness of fit will be ranked as a low rank can be avoided, and it becomes possible to search a desired document efficiently.

[0058]The suitable goodness-of-fit range can be easily specified by changing the goodness of fit of each document into the relative value over the maximum, and specifying the goodness-of-fit range specification in retrieval required with a relative value.

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]
[Drawing 1]The block diagram showing the composition of the document retrieval system in a 1st embodiment of this invention,
[Drawing 2]The flow chart showing the procedure of the retrieval processing in a 1st embodiment,
[Drawing 3]The figure showing an example of document data,
[Drawing 4]The figure showing an example of the retrieval-required character string in a 1st embodiment,
[Drawing 5]The figure showing an example of the contents of an output of the logic operation means in a 1st embodiment,
[Drawing 6]The figure showing an example of the contents of an output of the sorting information acquisition means in a 1st embodiment,
[Drawing 7]The figure showing an example of the contents of an output of the intermediate result cut-off point means in a 1st embodiment,
[Drawing 8]The figure the result in a 1st embodiment showing an example of the contents of an output of a rearrangement means,
[Drawing 9]The block diagram showing the composition of the document retrieval system in a 2nd embodiment of this invention,
[Drawing 10]The flow chart showing the procedure of the retrieval processing in a 2nd embodiment,
[Drawing 11]The figure showing an example of the contents of an output of the intermediate result division means in a 2nd embodiment,
[Drawing 12]The figure the result in a 2nd embodiment showing an example of the contents of an output of a rearrangement means,
[Drawing 13]The block diagram showing the composition of the conventional document

retrieval system,

[Drawing 14]The flow chart showing the procedure of the conventional retrieval processing,

[Drawing 15]The figure showing an example of a retrieval-required character string,

[Drawing 16]It is a figure showing an example which is the contents of an output of a result rearrangement means.

[Description of Notations]

101, 901, 1301 document data

102, 902, and 1302 Dictionary

103, 903, and 1303 Word frequency information extraction means

104, 904, and 1304 Word frequency index

105, 905, and 1305 Bibliographic-items extraction means

106, 906, and 1306 Bibliographic-items index

107, 907, and 1307 Retrieval-required input means

108, 908, and 1308 Word frequency calculation means

109, 909, and 1309 Logic operation means

110, 910, and 1310 Goodness-of-fit calculation means

111 and 911 Relative goodness-of-fit calculation means

112, 912, a 1311 sorting-information acquisition means

113 Intermediate result cut-off point means

114, 914, and 1312 Result rearrangement means

115, 915, and 1313 Result display means

913 Intermediate result division means

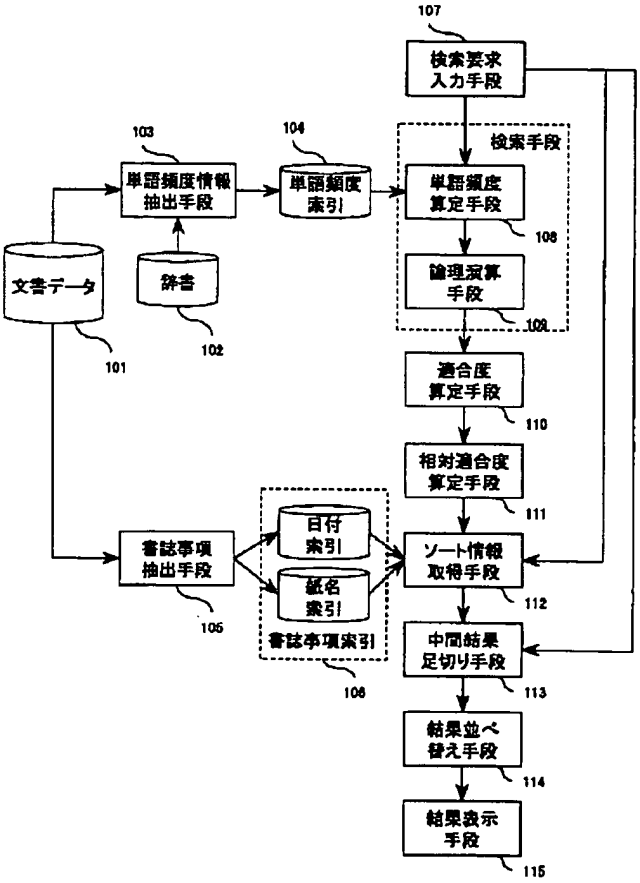
[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DRAWINGS

[Drawing 1]



[Drawing 3]

05^F19870317^FNY円急反落(海外外為)。十六日のニューヨーク外国為替市場の円相場は急反落。・・・^R02^F19870317^F少年のナイフ事件続発、・・・

[Drawing 4]

松下 AND 新製品 @HIDUKE @SHIMEI $70:

[Drawing 5]

レコード内部番号	「松下」のレコード内出現頻度	「新製品」のレコード内出現頻度
10	1	2
80	3	3
.	.	.
.	.	.
.	.	.
.	.	.
2,100,255	4	2
2,100,258	2	4
2,100,257	4	1
2,100,259	6	6
2,100,361	2	1
2,100,365	1	2

3289件

総レコード数	2,100,865
「松下」の出現レコード数	13,271
「新製品」の出現レコード数	8,280

[Drawing 6]

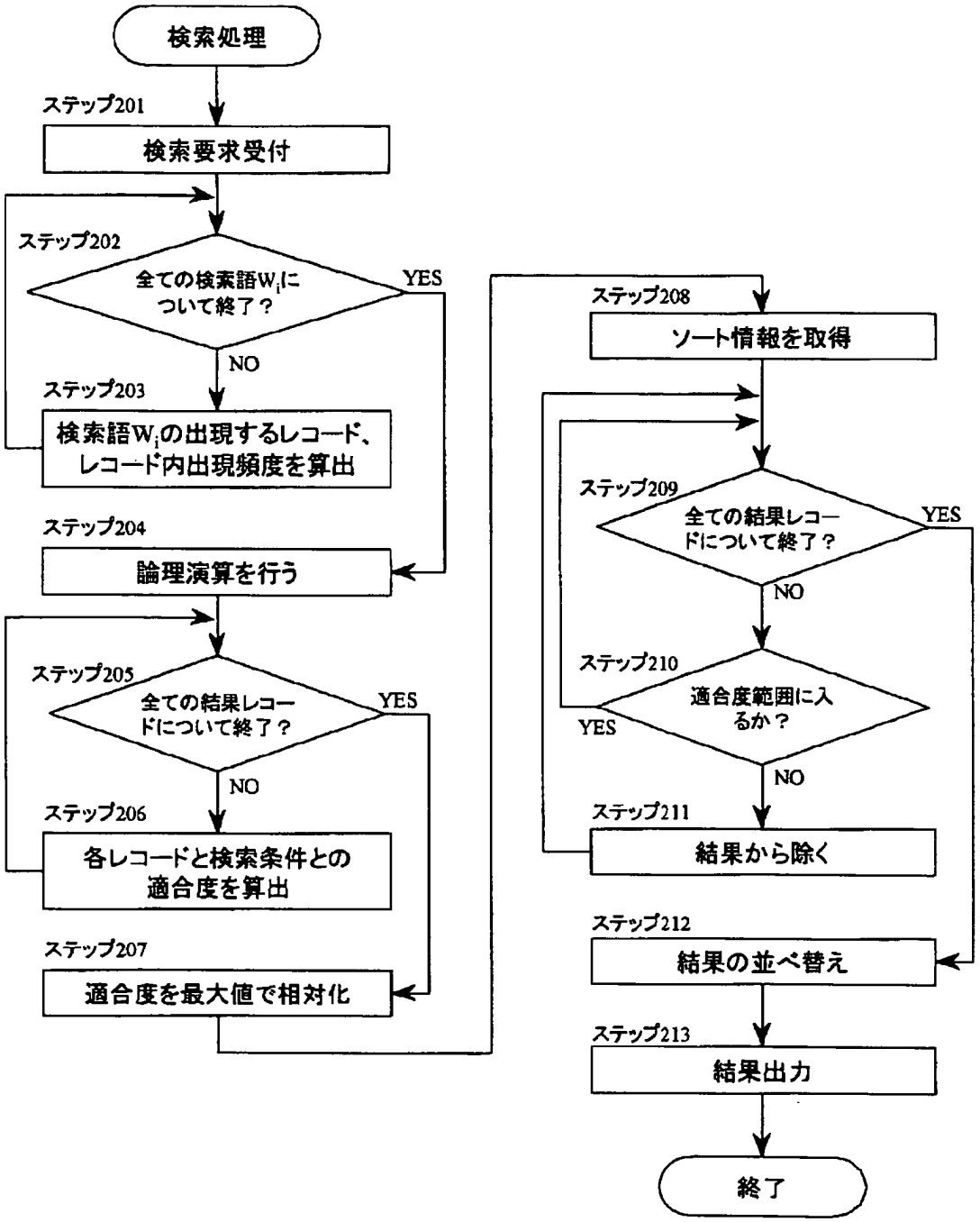
レコード内部番号	日付	紙名コード	適合度
10	19870502	02	84.1
80	19870502	01	50.2
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
2,100,255	19990725	05	77.5
2,100,258	19990725	05	70.6
2,100,257	19990725	05	100
2,100,259	19990725	05	1.5
2,100,361	19990725	01	88.8
2,100,365	19990725	01	5.3

3289件

[Drawing 15]

松下 AND 新製品 @HIDUKE @SHIMEI

[Drawing 2]



[Drawing 7]

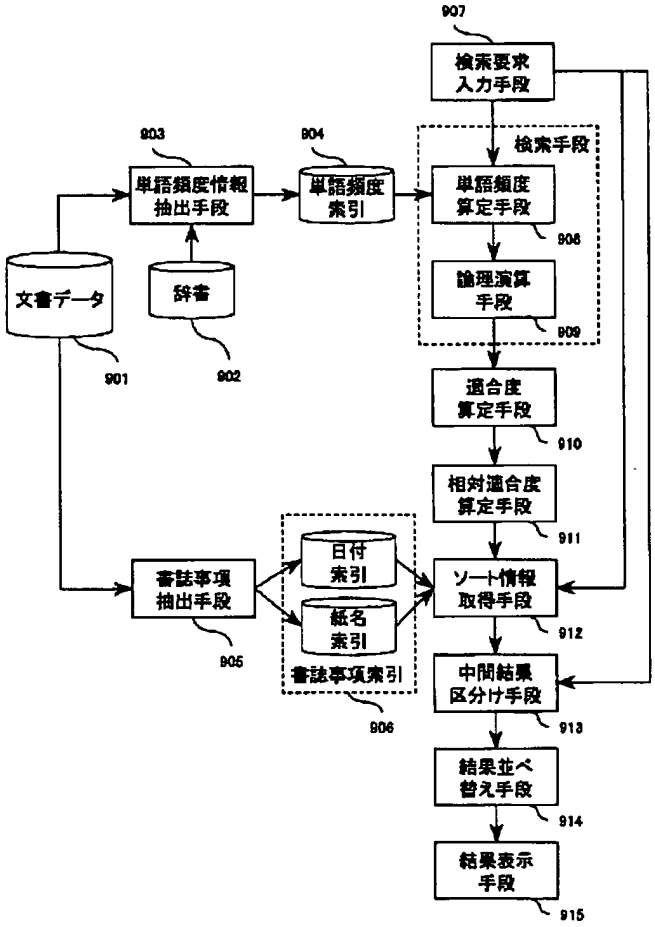
レコード内部番号	日付	組名コード	適合度
10	19870502	02	84.1
・	・	・	・
・	・	・	・
2,100,255	19990725	05	77.5
2,100,258	19990725	05	70.6
2,100,257	19990725	05	100
2,100,361	19990725	01	88.8

82件

[Drawing 8]

順位	レコード内部番号	日付	紙名コード	適合度
1	2,100,361	19990725	01	88.8
2	2,100,257	19990725	05	100
3	2,100,255	19990725	05	77.5
4	2,100,256	19990725	05	70.6
・	・	・	・	・
・	・	・	・	・
82	10	19870502	02	84.1

[Drawing 9]



[Drawing 11]

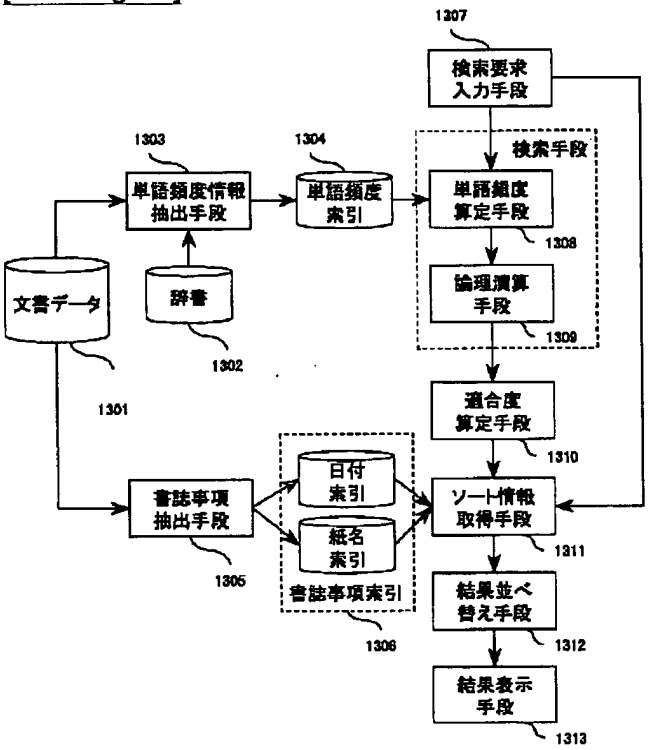
レコード内部番号	日付	紙名コード	適合度	区分けフラグ
10	19870502	02	84.1	1
80	19870502	01	50.2	2
・	・	・	・	
・	・	・	・	
・	・	・	・	
・	・	・	・	
2,100,255	19990725	05	77.5	1
2,100,256	19990725	05	70.6	1
2,100,257	19990725	05	100	1
2,100,259	19990726	05	1.5	2
2,100,361	19990726	01	88.8	1
2,100,365	19990725	01	5.3	2

3289件

[Drawing 12]

順位	レコード内部番号	日付	紙名コード	適合度	区分けフラグ
1	2,100,361	19990725	01	88.8	1
2	2,100,257	19990725	05	100	1
3	2,100,255	19990725	05	77.5	1
4	2,100,256	19990725	05	70.6	1
・	・	・	・	・	・
・	・	・	・	・	・
82	10	19870502	02	84.1	1
83	2,100,365	19990725	01	5.3	2
84	2,100,259	19990725	05	1.5	2
・	・	・	・	・	・
・	・	・	・	・	・
・	・	・	・	・	・
3289	80	19870502	01	50.2	2

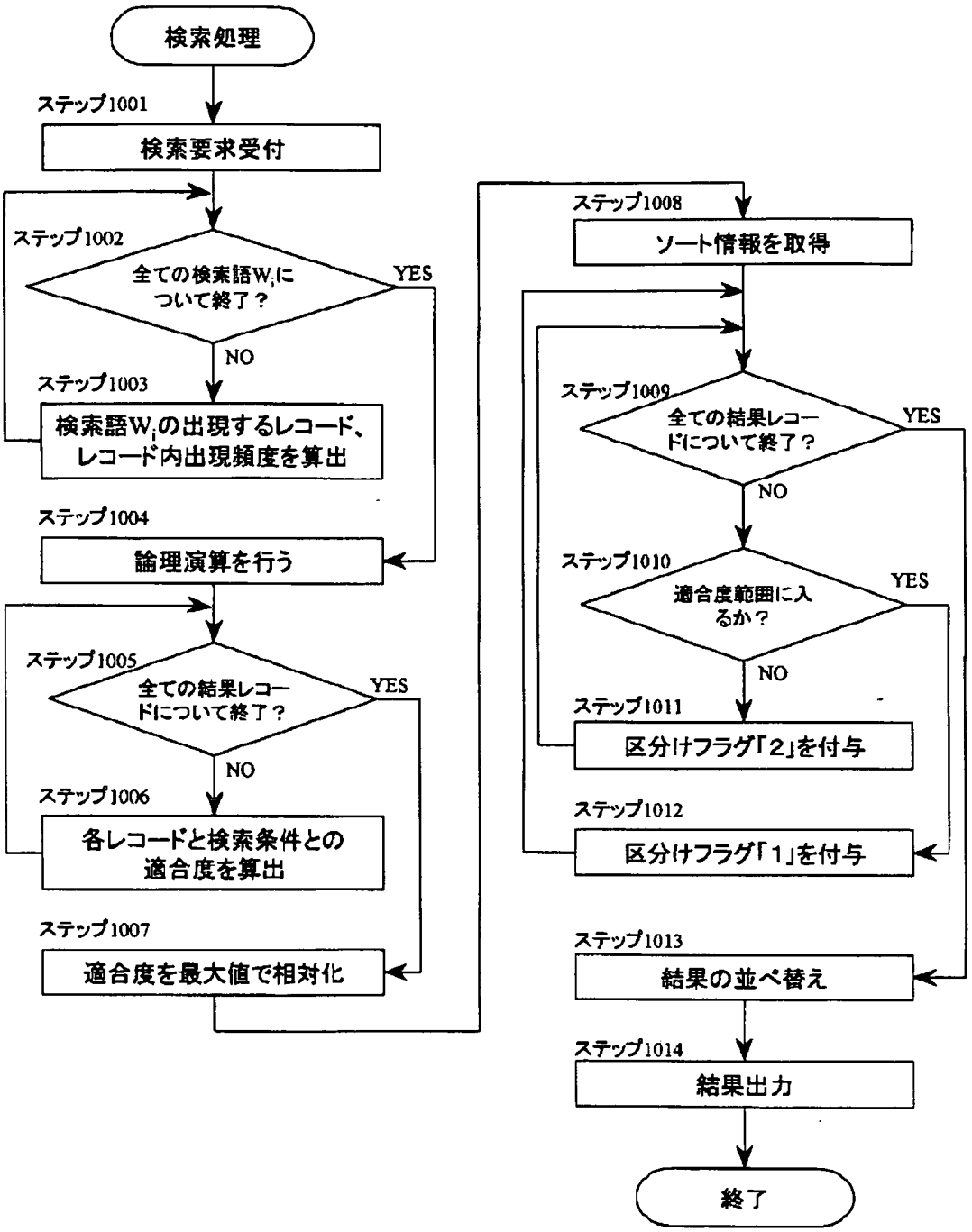
[Drawing 13]



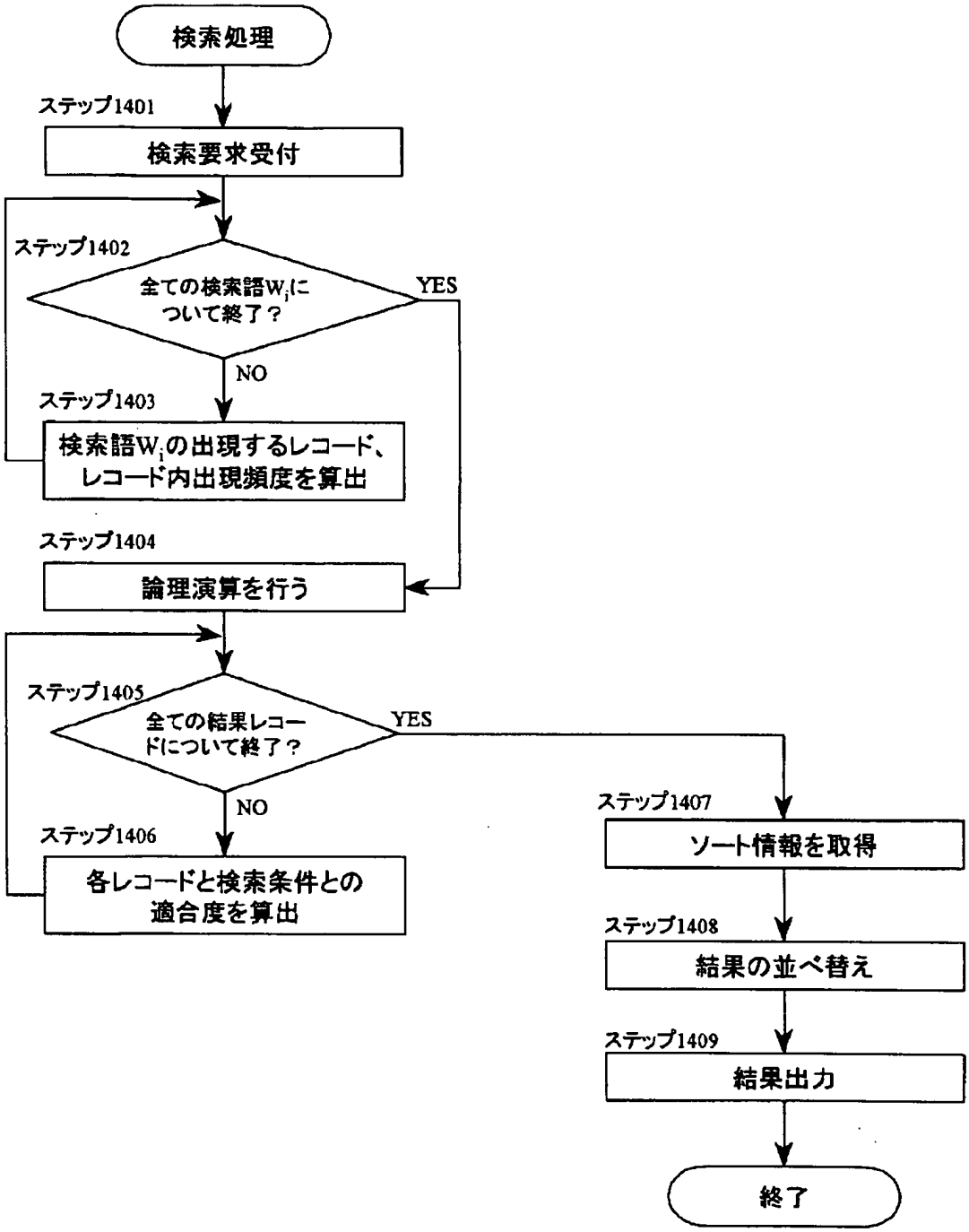
[Drawing 16]

順位	レコード内部番号	日付	紙名コード	適合度
1	2,100,361	19990725	01	8.88
2	2,100,365	19990725	01	0.53
3	2,100,257	19990725	05	10
4	2,100,255	19990725	05	7.75
5	2,100,256	19990725	05	7.06
6	2,100,259	19990725	05	0.15
・	・	・	・	・
・	・	・	・	・
・	・	・	・	・
・	・	・	・	・
3288	80	19870502	01	5.02
3289	10	19870502	02	8.41

[Drawing 10]


検索処理
ステップ1001
検索要求受付
ステップ1002
全ての検索語W_iについて終了？
YES
NO
ステップ1003
検索語W_iの出現するレコード、レコード内出現頻度を算出
ステップ1004
論理演算を行う
ステップ1005
全ての結果レコードについて終了？
YES
NO
ステップ1006
各レコードと検索条件との適合度を算出
ステップ1007
適合度を最大値で相対化
ステップ1008
ソート情報を取得
ステップ1009
全ての結果レコードについて終了？
YES
NO
ステップ1010
適合度範囲に入るか？
YES
NO
ステップ1011
区分けフラグ「2」を付与
ステップ1012
区分けフラグ「1」を付与
ステップ1013
結果の並べ替え
ステップ1014
結果出力
終了

[Drawing 14]



[Translation done.]